Analysis of trajectory data

1. Fundamental quantities
   Total energy, temperature, pressure, volume (or density)

2. Structural quantities
   Root Mean Square Deviation (RMSD)
   Distribution functions (e.g., pair and radial)
   Conformational analysis (e.g., Ramachandran plots, rotamers)

3. Dynamical quantities
   Time correlation functions & transport coefficients
   Free energy calculations using perturbation, umbrella sampling and steered MD methods
   Continuum electrostatics – Poisson and Poisson Boltzmann eq’s
   Protein interactions, ligand binding and docking methods
Fundamental quantities

- Total energy: strictly conserved but due to numerical errors it may drift.
- Temperature: would be constant in a macroscopic system (fluctuations are proportional to $1/\sqrt{N}$, hence negligible). But in a small system, they will fluctuate around a base line. Temperature coupling ensures that the base line does not drift from the set temperature.
- Pressure: similar to temperature, though fluctuations are much larger compared to the set value of 1 atm.
- Volume (or density): fixed in the NVT ensemble but varies in NPT. Therefore, it needs to be monitored during the initial equilibration to make sure that the system has converged to the correct volume (density). Relatively quick for globular proteins but could take a long time (~1 ns) for systems involving lipids (membrane proteins).
Root Mean Square Deviation

For an N-atom system, variance and RMSD at time t are defined as

\[
\sigma(t) = \frac{1}{N} \sum_{i=1}^{N} [r_i(t) - r_i(0)]^2, \quad \text{RMSD} = \sqrt{\sigma}
\]

Where \(r_i(0)\) are the reference coordinates. Usually they are taken from the first frame in an MD simulation, though they can also be taken from the PDB or a target structure. Because side chains in proteins are different, it is common to use a restricted set of atoms such as backbone or \(C_\alpha\). RMSD is very useful in monitoring the approach to equilibrium (typically, signaled by the appearance of a broad shoulder).

It is a good practice to keep monitoring RMSD during production runs to ensure that the system stays near equilibrium.
Evolution of the RMSD for the backbone atoms of ubiquitin
Distribution functions

Pair distribution function $g_{ij}(r)$ gives the probability of finding a pair of atoms $(i, j)$ at a distance $r$. It is obtained by sampling the distance $r_{ij}$ in MD simulations and placing each distance in an appropriate bin, $[r, r+\Delta r]$. Pair distribution functions are used in characterizing correlations between pair of atoms, e.g., hydrogen bonds, cation-carbonyl oxygen. The peak gives the average distance and the width, the strength of the interaction. When the distribution is isotropic (e.g., ion-water in an electrolyte), one samples all the atoms in the spherical shell, $[r, r+\Delta r]$. Thus to obtain the radial distribution function, the sampled number needs to be normalized by the volume $\sim 4\pi r^2 \Delta r$ (as well as density)

$$g_{iw}(r) = \frac{\langle \Delta N(r) \rangle}{4\pi r^2 \Delta r \rho}$$
Radial distribution functions exhibit oscillations, where each maxima is identified with a coordination number. They are obtained by integrating $g(r)$ between two neighbouring minima, e.g. denoting the first minimum by $r_{\text{min}}$, the first coordination number is given by

$$N_1 = \int_{0}^{r_{\text{min}}} 4\pi r^2 g(r) \, dr$$
Conformational analysis

In proteins, the bond lengths and bond angles are more or less fixed. Thus we are mainly interested in conformations of the torsional angles. As the shape of a protein is determined by the backbone atoms, the torsional angles, $\phi$ (N--C$_{\alpha}$) and $\psi$ (C--C$_{\alpha}$), are of particular interest. These are conveniently analyzed using the Ramachandran plots. Conformational changes in a protein during MD simulations can be most easily revealed by plotting these torsional angles as a function of time.

Also of interest are the torsional angles of the side chains, $\chi_1$, $\chi_2$, etc. Each side chain has several energy minima (called rotamers), which are separated by low-energy barriers ($\sim kT$). Thus transitions between different rotamers is readily achievable, and they could play functional roles, especially for charged side chains.
Time correlation functions & transport coefficients

At room temperature, all the atoms in the simulation system are in a constant motion characterized by their average kinetic energy: \((3/2)kT\). Free atoms or molecules diffuse according to the relationship

\[
\langle \mathbf{r}^2(t) \rangle = 6Dt
\]

While this relationship can be used to determine the diffusion coefficient, more robust results can be obtained using correlation functions, e.g.,

\(D\) is related to the velocity autocorrelation function as

\[
D_v = \frac{1}{3} \int_0^\infty \langle \mathbf{v}_v(0) \cdot \mathbf{v}_v(t) \rangle dt
\]

Similarly, conductance of charged particles is given by the Kubo formula

\[
\sigma_v = \frac{1}{3VkT} \int_0^\infty \langle \mathbf{J}_v(0) \cdot \mathbf{J}_v(t) \rangle dt, \quad \mathbf{J}_v = \sum_i q_{vi} \mathbf{v}_i \quad \text{(current)}
\]
Bound atoms or groups of atoms fluctuate around a mean value. Most of the high-frequency fluctuations (e.g. H atoms), do not directly contribute to the protein function. Nevertheless they serve as “lubricant” that enables large scale motions in proteins (e.g. domain motions) that do play a significant role in their function. As mentioned earlier, large scale motions occur in a ~ ms to s time scale, hence are beyond the present MD simulations. (Normal mode analysis provides an alternative.) But torsional fluctuations occur in the ns time domain, and can be studied in MD simulations using time correlation functions, e.g.

$$C_\phi(t) = \frac{\langle \Delta \phi(0) \cdot \Delta \phi(t) \rangle}{\langle |\Delta \phi(0)|^2 \rangle}$$

These typically decay exponentially and such fluctuations can be described using the Langevin equation.
Free energy calculations

Free energy is the most important quantity that characterizes a dynamical process. Calculation of the absolute free energies is difficult in MD simulations. However free energy differences can be estimated using several methods. The most common method is involved making an alchemical transformation from an initial state described by a Hamiltonian $H_0$ to a final one described by $H_1$. To ensure the convergence of the results, this transformation is done gradually using a parameter $\lambda$

$$H(\lambda) = (1 - \lambda)H_0 + \lambda H_1$$

Thermodynamic integration:

$$\Delta G = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle d\lambda$$

This integral is evaluated most efficiently using a Gaussian quadrature.
Free energy perturbation: Here the interval between $\lambda = 0$ and $1$ is divided into $n$ subintervals with $\{\lambda_i, i = 0, n\}$. For each $\lambda_i$ value, the free energy difference is calculated from the ensemble average

$$\Delta G(\lambda_i \rightarrow \lambda_{i+1}) = -kT \ln \langle \exp[-(H(\lambda_{i+1}) - H(\lambda_i))] / kT \rangle_{\lambda_i}$$

The total free energy change is then obtained by summing the contributions from each subinterval

$$\Delta G = \sum_{i=0}^{n-1} \Delta G(\lambda_i \rightarrow \lambda_{i+1})$$

The number of subintervals is chosen such that the free energy change at each step is at most $2\ kT$, otherwise the method loses its validity.
Umbrella sampling:

In this method one samples the densities along a reaction coordinate and determines the potential of mean force (PMF) from the Boltzmann eq.

\[
\langle \rho(z) \rangle = \langle \rho(z_0) \rangle e^{[W(z) - W(z_0)]/kT} \rightarrow W(z) = W(z_0) - kT \ln \left( \frac{\langle \rho(z) \rangle}{\langle \rho(z_0) \rangle} \right)
\]

Here \( z_0 \) is a reference point, e.g. a point in bulk where \( W \) vanishes. In general, a particle cannot be adequately sampled at high-energy points. To counter that, one introduces harmonic potentials, which restrain the particle at desired points, and then unbias its effect. For convenience, one introduces umbrella potential at regular intervals along the reaction coordinate (e.g. \( \sim 0.5 \) Å). The PMF’s obtained in each interval is are unbiased and optimally combined using the Weighted Histogram Analysis Method (WHAM).